Can students evaluate their understanding of cause-and-effect relations? The effects of diagram completion on monitoring accuracy

Acta Psychologica 151 (2014) 143–154

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/ locate/actpsy

Can students evaluate their understanding of cause-and-effect relations? The effects of diagram completion on monitoring accuracy

Mariëtte H. van Loon [a,*], Anique B.H. de Bruin [a], Tamara van Gog [b], Jeroen J.G. van Merriënboer [a], John Dunlosky [c]

[a] Maastricht University, Department of Educational Development & Research and Graduate School of Health Professions Education, P.O. Box 616, 6200 MD Maastricht, The Netherlands
[b] Erasmus University Rotterdam, Department of Psychology, P.O. Box 1783, 3000 DR Rotterdam, The Netherlands
[c] Kent State University, Department of Psychology, P.O. Box 5190, Kent, OH 44242-0001, USA

ARTICLE INFO

Article history:
Received 27 January 2014
Received in revised form 3 June 2014
Accepted 10 June 2014
Available online 28 June 2014

PsycINFO codes:
2343
3500

Keywords:
Monitoring
Regulation
Diagram
Causal relations
Adolescents

ABSTRACT

For effective self-regulated study of expository texts, it is crucial that learners can accurately monitor their understanding of cause-and-effect relations. This study aimed to improve adolescents' monitoring accuracy using a diagram completion task. Participants read six texts, predicted performance, selected texts for restudy, and were tested for comprehension. Three groups were compared, in which learners either completed causal diagrams immediately after reading, completed them after a delay, or received no-diagram control instructions. Accuracy of predictions of performance was highest for learning of causal relations following delayed diagram completion. Completing delayed diagrams focused learners specifically on their learning of causal relations, so this task did not improve monitoring of learning of factual information. When selecting texts for restudy, the participants followed their predictions of performance to the same degree, regardless of monitoring accuracy. Fine-grained analyses also showed that, when completing delayed diagrams, learners based judgments on diagnostic cues that indicated actual understanding of connections between events in the text. Most important, delayed diagram completion can improve adolescents' ability to monitor their learning of cause-and-effect relations.

1. Introduction

The ability to read and comprehend expository texts is crucial for adolescents' school progress and future careers (Otero, Leon, & Graesser, 2002). Integral to the process of text comprehension is one's ability to monitor whether the information that is read has been understood and can be remembered (Nelson & Narens, 1990). Based on the output from monitoring, students decide how well they have achieved their learning goals, and then regulate their learning by further reading passages or dropping them from study (De Bruin & Van Gog, 2012; Koriat, 2012). Readers who can accurately monitor their current level of understanding are able to learn more from textual information (Dunlosky & Rawson, 2012; Thiede, Anderson, & Therriault, 2003), presumably because they can strategically decide which passages are not well learned and need further study (Son & Metcalfe, 2000; Thiede et al., 2003). Some interventions have been identified that can improve monitoring accuracy when learning from texts, such as generating keywords (Thiede et al., 2003), summaries (Thiede & Anderson, 2003), or sentences (Van Loon, De Bruin, Van Gog & Van Merriënboer, 2013a). However, little is known about how to improve monitoring accuracy when participants are tested for higher-order understanding of complex materials (e.g., Ackerman, Leiser, & Shpigelman, 2013).

It is especially interesting to study adolescents' metacognitive skills in an educational context, because adolescents are encountering the complex demands of becoming self-regulated learners (Alvermann, 2002). Even though adolescents' metacognitive skills are more developed than those of primary-school learners, these skills seem not yet to be as developed as those of adults (Koriat, Ackerman, Adiv, Lockl, & Schneider, 2014). Young adults have been demonstrated to struggle when monitoring their learning of complex information (e.g. Ozuru, Kurby, & McNamara, 2012; Thiede et al., 2003). Adolescents may encounter even more difficulties when monitoring and regulating learning, because their working memory capacity and processing speed are still developing until adulthood (Fry & Hale, 1996; Kail & Salthouse, 1994). It has been proposed that developmental factors related to learners' cognitive capacity play an important role when monitoring learning of difficult tasks (Krebs & Roebers, 2010, 2011; Roebers, Von der Linden, & Howie, 2007). However, research addressing monitoring accuracy during complex learning was mainly conducted with college students in laboratory settings. Even though this research provided valuable theoretical knowledge, important questions remain about how findings can be applied for younger learners in education. In particular, evidence is scarce on how to improve adolescents' metacognitive

* Corresponding author. Tel.: +31 43 3885720; fax: +31 43 3885779.
E-mail address: m.vanloon@maastrichtuniversity.nl (M.H. van Loon).

http://dx.doi.org/10.1016/j.actpsy.2014.06.007
0001-6918/

monitoring and regulation when studying complex texts. In the present study, we presumed that appropriate support (with pre-structured, to-be-completed diagrams) would help them to accurately monitor their learning. Therefore, the main goal of the current study is to improve the accuracy of adolescents' monitoring judgments when reading complex expository texts containing causal relations.

To motivate our study, we first briefly discuss students' difficulties in evaluating their comprehension of text and how it can be improved. Based on theory of text comprehension, we then argue that supplementing expository texts with to-be-completed diagrams will help students more accurately evaluate their comprehension. Finally, we discuss the importance of having students complete delayed diagrams, and we then summarize our theoretical predictions.

1.1. Monitoring accuracy

Readers' monitoring of their text comprehension is often inaccurate, as demonstrated, for instance, by De Bruin, Thiede, Camp, and Redford (2011). They asked participants to read a set of texts. After reading, the participants predicted for each of the texts how well they would do on a future comprehension test. Monitoring accuracy was operationalized as the correlation between predictions and test performance (Nelson, 1984). When students were not supported by a metacognitive prompt that required them to generate keywords, their monitoring accuracy was not significantly higher than zero. This outcome indicates that readers were unable to accurately monitor their text comprehension. Unfortunately, this low level of monitoring accuracy is commonly reported throughout the literature. In their review of research over the past two decades, Thiede, Griffin, Wiley, and Redford (2009) reported that the average correlation across 57 studies was only .27 (see also Dunlosky & Lipko, 2007; Maki, 1998).

Despite this poor monitoring accuracy, research on text comprehension suggests that it can be improved if readers are encouraged to evaluate their understanding of the gist of a studied text (Anderson & Thiede, 2008; De Bruin et al., 2011). Knowledge about text can be represented at different levels (Graesser, Singer, & Trabasso, 1994; Kintsch, 1998; Van den Broek, Rapp, & Kendeou, 2005). For complete text comprehension, learners must go beyond text base processing of factual information and need to establish a coherent representation of the gist of the text, usually termed the situation model (Kintsch, 1998). To achieve this level of gist comprehension, learners need to generate inferences by connecting relations between elements presented in the text. Thus, to improve accuracy when monitoring understanding of a text, learners need to base their predictions of performance on cues that arise from processing information about the gist of a text (for a detailed discussion, see Rawson, Dunlosky, & Thiede, 2000). Doing so should provide them with cues that are more diagnostic of their subsequent comprehension test performance; hence, using these diagnostic cues to predict future performance should lead to higher levels of monitoring accuracy (e.g., Brunswik, 1956; Koriat, 1997).

1.2. Improving monitoring accuracy when learning causal relations

With respect to understanding of expository texts, gist comprehension depends to a large extent on a reader's ability to connect and understand the cause-and-effect relations in the text (Graesser et al., 1994). To support learners with comprehension of complex texts, some textbooks depict the structure of cause-and-effect relations with causal diagrams, wherein events presented in the text are connected with arrows (Cromley, Snyder-Hogan, & Luciw-Dubas, 2010; Cromley et al., 2013; McCrudden, Magliano, & Schraw, 2011; McCrudden, Schraw, Lehman, & Poliquin, 2007). Providing diagrams in addition to the expository text has been demonstrated to improve gist comprehension (Butcher, 2006; Mayer, 2003). Interestingly, besides presenting completed diagrams as an aid to foster learning, instructing learners to self-generate diagrams has been shown to be promising to foster comprehension. When learners are asked to depict cause-and-effect relations during reading by drawing causal diagrams, text understanding improves in comparison to when they were asked to re-read or summarize the text (Gobert & Clement, 1999). Presumably, generating diagrams promoted deeper processing and resulted in a richer mental model of the information presented in the text (Gobert & Clement, 1999). Note that in these studies, diagram presentation and generation were used as a means to promote deeper processing of information during text study. However, during self-regulated learning, emphasis is placed on improving *monitoring* of comprehension, because study decisions, and thus learning efforts, are based on monitoring judgments (Metcalfe & Finn, 2008; Thiede et al., 2003). Research by Thiede et al. (2003) showed that instructions that improve monitoring accuracy and study selections led to better final text comprehension scores in a self-regulated learning setting than instructions aiming to improve deeper text processing during initial text study.

In the present study, we aimed to improve monitoring accuracy and subsequent restudy selections with the use of diagram completion instructions. We hypothesized that, in addition to improve deeper text processing, asking learners to generate causal diagrams might also be a promising instructional strategy to improve monitoring accuracy for learning of causal relations in text. To help learners focus on cues diagnostic of comprehension of causal relations, we developed a diagram completion task in which readers attempted to depict the steps in a causal chain presented in the text (see Fig. 1). To support learners with this task, the structure of the causal diagram was presented to them, and they were instructed to complete the diagram by filling out the empty text boxes. When completing the diagram about a studied text, the learner must identify the steps presented in the text and infer the relations between these steps. In the present context, completing the diagrams is expected to provide diagnostic cues that indicate whether readers have understood the cause-and-effect relations within the text and hence to boost the accuracy of their monitoring judgments. More generally, this theoretical claim is based on the cue-utilization framework (e.g., Koriat, 1997), which states that to monitor their learning, people use cues that are accessed prior to making a judgment, so that the accuracy of the judgments will be determined by how well those cues predict criterion test performance (called *cue diagnosticity*, which is described further below). When studying for deep comprehension of higher-order ideas in a text, diagram completion is expected to focus learners' attention on cues that are diagnostic of future test performance by indicating whether or not they are able to connect the events that were presented in the text.

1.3. Timing of the diagram completion task

Prior research suggests that any benefits of using diagrams to improve monitoring accuracy might be greatest when learners complete the diagrams some time after studying a text, rather than immediately after studying it. For instance, Thiede and Anderson (2003) showed that monitoring accuracy could be improved with the use of generative self-tests, such as summary or keyword generation. These tasks presumably focus readers' attention on diagnostic cues related to their comprehension of the gist of texts. When diagnostic cues are used to inform learners' predictions, they can support more accurate monitoring judgments (Redford, Thiede, Wiley, & Griffin, 2012; Thiede & Anderson, 2003). Despite the improvements in monitoring accuracy, these generation tasks only had beneficial effects on monitoring when both tasks and predictions were made after a delay, indicating the importance of the timing of the generation tasks. Because learners still have the text base level available when performing a generative self-test immediately after study, this task does not inform them about their long-term memory for the gist of the text (Anderson & Thiede, 2008; Kintsch, Welsch, Schmalhofer, & Zimny, 1990). By contrast, when learners perform a generation task after a short delay, typically only a few minutes after reading, they would need to access the gist

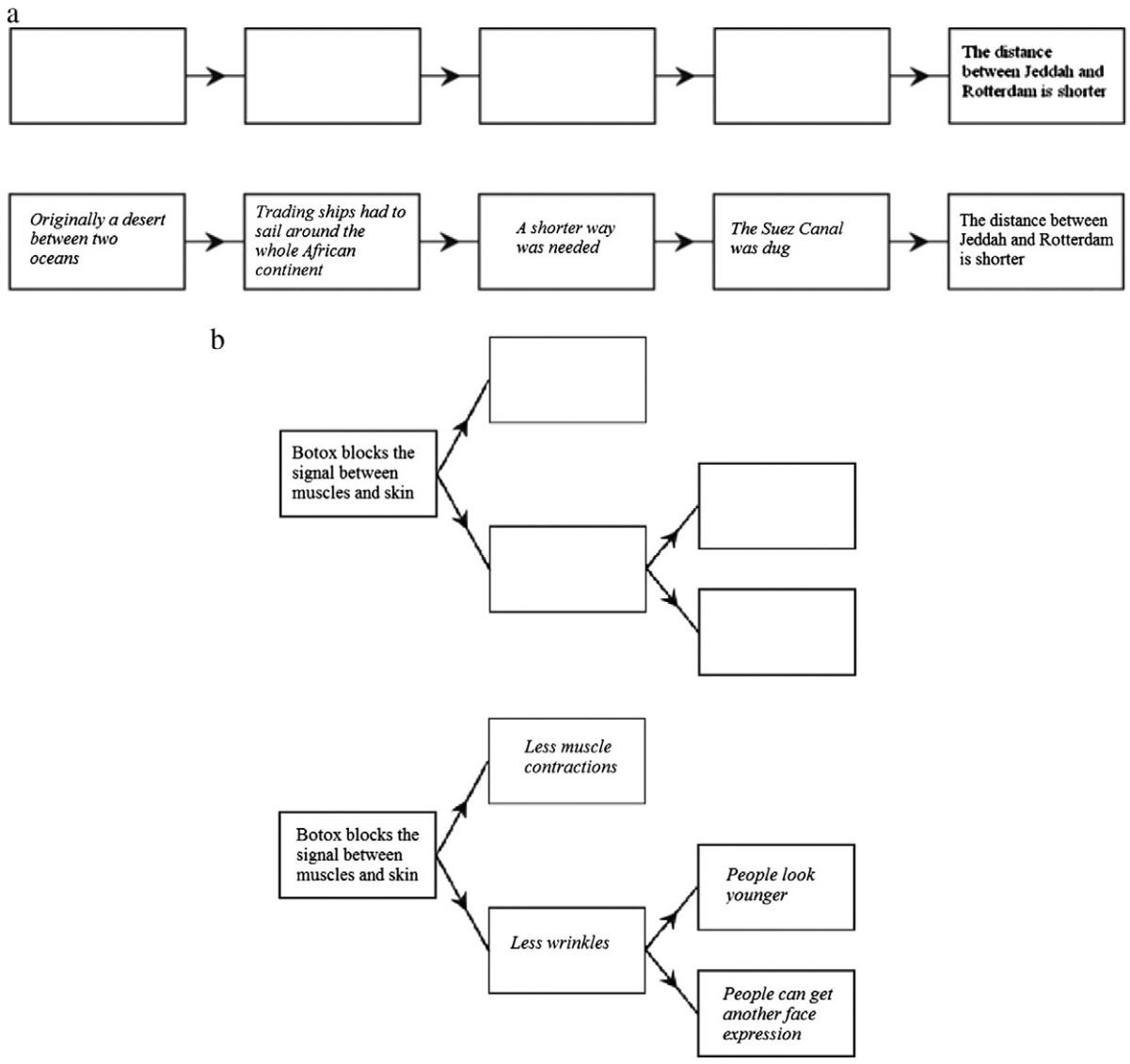

Fig. 1. a. An empty and a correctly completed diagram for the text 'Suez Canal'. b. An empty and a correctly completed diagram for the text 'Botox'.

representation of the text from their long-term memory (Thiede, Dunlosky, Griffin, & Wiley, 2005). Thiede et al. (2005) demonstrated that the delay between reading and the generation task is the only lag which is critical to improvement of monitoring accuracy. Presumably, delaying the generation and the judgment tasks provides learners with better diagnostic cues about their understanding of the text, and using these cues improves monitoring accuracy (Thiede et al., 2009). Based on this evidence, we presume that the timing of the diagram completion task is an important factor when aiming to improve monitoring accuracy, so that completing the diagrams some time after reading texts has more beneficial effects on accuracy than completing diagrams immediately after reading.

1.4. Present study

A challenge for research on metacognition is to improve monitoring accuracy for understanding of complex study materials (e.g., Ackerman et al., 2013). We aimed to investigate the effects of asking adolescents to complete causal diagrams on monitoring accuracy for understanding of cause-and-effect relations. Three groups were compared; one group completed diagrams immediately after reading texts; one group completed these after a short delay, and a no-diagram group that did not complete diagrams (but just made delayed predictions) was also included. All participants made two predictions of performance (POPs) for each text: One in which they predicted how well they would perform on a test about causal relations, and one in which they predicted how well they would perform on a test about factual information (cf. Ackerman & Goldsmith, 2011). Note that for all groups, the POPs were made at a delay after reading the texts.

We hypothesize that (a) monitoring accuracy for learning of causal relations will be greater when participants complete diagrams than when they do not (Hypothesis 1); (b) monitoring for learning of causal relations will be most accurate when diagrams are completed after a delay (Hypothesis 2), and (c) there will be no differences across groups in monitoring accuracy for learning of factual information, because the diagram completion task is expected to focus learners explicitly on their learning of causal relations rather than factual information (Hypothesis 3).

Learners may use a variety of cues when monitoring their learning; for instance, they may judge that they understand one text better than another because they could generate more relations for one text than another. As introduced above, learners' POPs will be accurate to the degree that the cues they use to make the predictions are indicative of actual learning, which is referred to as cue diagnosticity (see Brunswik, 1956, who referred to this construct as ecological validity). Cue diagnosticity is measured using a within-participant correlation

between the cue and test performance. In this example, the cue of generating relations (where more are generated for some texts than other texts) would be diagnostic if it positively correlated with test performance across texts. Based on this framework, the current experimental design allowed us to explore why diagrams improve monitoring accuracy, because the quality of learners' diagrams can be scored and used to estimate the diagnosticity of available cues (Question 1a). For instance, if the number of correctly completed causal relations for each diagram (a cue) correlates highly with future test performance, this cue would be considered highly diagnostic. We can also evaluate utilization of these cues by investigating the degree to which learners' monitoring judgments are related to this diagnostic cue and hence can benefit from it (Question 1b). *Cue utilization* is measured using a within-participant correlation between the diagram completion cues and POPs. In summary, to isolate the causes of prediction accuracy, we also estimated the diagnosticity and utilization of a variety of cues that arise from diagram completion.

Finally, although our primary goal was to evaluate whether delayed diagrams would improve monitoring accuracy, accurate POPs can improve comprehension only if a learner is given the opportunity to use monitoring (Thiede, 1999). When readers identify which information they have not yet correctly understood, doing so can guide them to select materials that need further study. Monitoring judgments have been demonstrated to be directly related to learners' selections for restudy (Dunlosky & Ariel, 2011; Metcalfe & Finn, 2008; Son & Metcalfe, 2000; Thiede et al., 2003). Hence, inaccurate monitoring can be a serious problem for learning, because if learners cannot accurately judge their level of comprehension, they cannot effectively decide which texts need further study. Therefore, a secondary goal of this study was to explore how readers used their predictions to make decisions about restudying. In particular, we expected a strong relation between predictions and restudy selections for all three groups, indicating that regardless of monitoring accuracy, regulation of study is based on people's judgments of how well they have learned the texts (Hypothesis 4).

2. Method

2.1. Participants

Participants were 123 adolescents (Mean age = 14.78, *SD* = .67, range 14–16 years) from the ninth grade of a secondary school in the south of The Netherlands. All participants followed secondary education in one of the two programs that lead to higher education: Seventy-three were in the third year of pre-university secondary education level (VWO; 6 year duration, highest level of secondary education), and 50 were in the third year of higher general secondary education level (HAVO; 5 year duration, middle level of secondary education). All participants were native Dutch speakers. The participants were tested in their classroom setting; each participant was assigned randomly to one of three groups in a computer classroom. Because not all computers were occupied in each session, this resulted in a slightly different number of participants in each group: immediate diagram completion, $n = 40$; delayed diagram completion, $n = 44$; no-diagram group, $n = 39$.

2.2. Materials and design

The experiment included five phases: text study, experimental task, POPs, restudy selections, and test (see Table 1).

2.2.1. Pre-reading Instructions

Before the experimental task started, the experimenter provided all participants with pre-reading instructions, to get accustomed to the types of texts, the distinction between causal relations and factual information, the POPs, and the test format. The participants were asked to read two example texts: "The Heart" and "Suburbs". To explain the difference between causal relations and factual information, after reading the example texts, they were provided with two examples of test questions about causal relations and two examples of test questions about factual information. Then, they were shown that the causal relations between the events in the text could be visualized by drawing a diagram. They were explained that causal events could occur in a serial manner, meaning that these follow each other in time, or that events can occur simultaneously (refer to Fig. 1 for examples of serial and simultaneous relations).

Then participants were presented with two blank diagrams on a flipover, one example diagram for the text "The Heart" (containing only serial relations), and one diagram text: "Suburbs" (containing both serial and simultaneous relations). The experimenter and the participants jointly completed the two diagrams, the experimenter ensured that everyone understood how to complete diagrams about cause-and-effect relations.

Table 1
Overview of the procedure.

Instructions		
Immediate-diagram	Delayed-diagram	No-diagram
Read text 1	Read text 1	Read text 1
Diagram text 1	Read text 2	Read text 2
Read text 2		
Diagram text 2		
	Read text 6	Read text 6
	Diagram text 1	Control task text 1
	Diagram text 2	Control task text 2
Read text 6		
Diagram text 6	Diagram text 6	Control task text 6
Prediction of performance causal relations text 1		
Prediction of performance factual information text 1		
Prediction of performance causal relations text 6		
Prediction of performance factual information text 6		
Indication which texts need restudy (no actual restudy)		
Test of causal relations text 1		
Test of factual information text 1		
Test of causal relations text 6		
Test of factual information text 6		

2.2.2. Text study

To select texts, a pilot study with 20 texts was conducted. Forty-one ninth-grade students read five texts, completed diagrams, provided POPs, and took a test. Based on this pilot study, six expository texts were selected for the experimental task. The pilot study established that the selected texts had a sufficient level of difficulty; mean performance on questions about causal relations was 40.83% (*SD* = 24.34), mean performance on questions about factual information was 34.17% (*SD* = 23.81). None of the participants exhibited ceiling or floor effects when tested for learning of causal relations and details. Furthermore, the pilot study showed that the participants were able to fill out diagram text boxes about the texts; mean filled out text boxes = 82.5% (*SD* = 24.05).

Causal relations can occur both in a serial and a simultaneous format (see Appendix A and Fig. 1 for an example of a text and a diagram containing serial relations and a text and a diagram containing both serial and simultaneous relations). To have some variety in the type of causal events in the texts, the texts were selected so that three of those contained only serial causal relations, and three texts contained both serial and simultaneous relations. The pilot study showed that these two text types had a comparable level of difficulty. Topics of the selected texts were: "Botox", "Sinking of metro cars", "Concrete constructions", "Money does not bring happiness", "The Suez Canal", and "Music makes smart" (see Appendix A for examples of the texts). The mean number of words per text was 171.27; range = 162–189 words. Texts were written so that they comprised five clauses to convey causal relations. All texts were presented in a single-paragraph format. The participants could read each text only once; they were not able to return to the text in another stage in the experiment.

2.2.3. Experimental task

When logging-in, participants were randomly assigned, by using the log-in number, to one of three experimental groups: 1) immediate diagram; 2) delayed diagram; and 3) no-diagram group. In the two diagram completion groups, the participants were provided with a diagram on the screen. The diagram textboxes and arrows were pre-structured, containing either a serial or a simultaneous cause-and-effect relation. Each diagram contained five text boxes which were connected with arrows. One of the text boxes was already filled; the participants were instructed to complete the diagram by typing a response in the other four text boxes. The participants could only continue to the next screen if they had typed a response in all four of the text boxes, and they could not return to the text when filling out the diagrams.

In the immediate diagram group, the participants completed a diagram immediately after reading each text. When they pressed the 'continue' button after reading the text, the following screen was a diagram for that text. In the delayed diagram group, the participants first read all six texts and then completed the six diagrams for these texts (see Thiede et al., 2003, for the same procedure used with the keyword generation strategy).

In the no-diagram group, the participants read all six texts and were then presented with a picture-matching task for which they had to match two pictures related to the topics of the read texts. The participants had to provide a response about the four differences between the pictures, by typing their response in the four text boxes below the pictures. To present the participants in the no-diagram group with the same information as the participants in the two diagram completion groups, above the pictures, they were provided with the same statement that was presented in the filled-out text box in the diagram completion groups.

2.2.4. POPs

For each text, the participants provided two POPs: one for their learning of causal relations, and one for their learning of factual information. When making POPs, the participants saw the title of the text on the screen, accompanied by two questions. The first POP question was: How many questions do you expect to complete correctly when tested for understanding of causal relations in this text? The second POP question was: How many questions do you expect to complete correctly when tested for factual information about this text? The POPs were provided with a mouse-click on a 6-point scale ranging from 0% to 100% (points on the scale matched 0%, 20%, 40%, 60%, 80%, and 100%). Note that these POPs are delayed after study and diagram completion, we decided to not include immediate predictions because they typically show poor accuracy (Nelson & Dunlosky, 1991; Van Loon, De Bruin, Van Gog, & Van Merriënboer, 2013a).

2.2.5. Restudy selections

Participants indicated which texts they would want to restudy by clicking on a grid with a 3×2 array in which each cell was filled by a title of a previously studied text (in line with the design by Thiede & Dunlosky, 1999). The order of the texts in this grid was randomized for every participant; zero to six texts could be selected for restudy. Note, however, that learners did not actually restudy any of the texts, because restudying the texts after POPs were made could inadvertently influence the relation between POPs and comprehension (Kimball & Metcalfe, 2003), which is the focus of the present research.

2.2.6. Test

The test consisted of six questions about causal relations (one for each text), and 30 questions about factual information (five per text). One causal relation question was always followed by five factual information questions on a text. The questions about facts required short answers. Examples of questions about causal relations and factual information are provided in Appendix B.

All tasks were presented in Dutch, and the order of the six text topics was randomized anew in each phase. Note that in the immediate diagram phase, the order of the text study was randomized, but the experimental diagram completion task had the same topic as the text that was previously read. Presentation of all materials was self-paced.

2.3. Procedure

The participants were tested in a computer room in their school. The number of participants per session ranged from 22 to 28. They completed the task in one session that lasted for approximately 1 h, and all tasks were self-paced. Table 1 depicts the procedure of the study.

Before the study task started, the participants received the pre-reading instructions. They were told that they would study six texts for a later test with questions on factual information and causal relations. Before study, they were presented with the two example texts and example test questions. They practiced with two questions about causal relations and four questions about factual information. Note that the students were not informed about the exact number of test questions they would be asked to answer at the later test. After practicing with the texts and test questions, they practiced with completing example diagrams. Then the participants were told that they would be asked to judge their learning by predicting future test performance, and they were provided with an example of the POP scale for causal relations and the POP scale for questions about factual information.

Following these classroom instructions, the participants used the log-on information presented on a sheet of paper next to their computer to start the text study task. They were not provided with feedback on their responses in any of the phases of the experiment. After studying each text, they pressed a key, and the page was removed and replaced by the next page. In the immediate diagram group, the participants completed a diagram about each text immediately after reading. In the delayed diagram group, the participants read all six texts and then completed all six diagrams. The no-diagram group read all six texts and then completed the picture-matching task. In all groups, the participants typed in their answers before being presented with the next page. They were instructed to type a "?" in a text box when they were not able to come up with a response.

For all participants, all predictions were made after a delay, i.e., POPs were made after reading the texts and performing the experimental task. After selecting texts for restudy, the participants received the instruction on the screen that they would not actually get to restudy those texts. When taking the test, the participants were shown the title of the texts, accompanied by the questions. All test questions were answered by typing in the answers on the computer.

2.4. Scoring of responses

2.4.1. Test performance

Responses on the questions about causal relations were scored as per McCrudden et al. (2011). Scores on these questions ranged from 0 to 4; the score refers to the amount of correctly stated causal relations. Comprehension was emphasized; therefore, responses were also scored as correct when the participants did not respond with what was literally stated in the text but instead responded with a response indicating that they understood what was implied or meant with the original text, i.e., a response indicating gist understanding. For example, the following test response about the text on Botox was scored as containing two correct relations: "The muscles in the skin relax (correct relation) because something is injected, that's why some facial expressions are not visible anymore (correct relation)". Two independent raters scored 25.1% (a total of 185) of all test responses on questions about causal relations (responses were scored using an ordinal scale, inter-rater agreement was high, intraclass correlation for reliability ratings of first rater = .80; average reliability = .89).

Responses on the questions about factual information were scored as omission, commission error, partially correct, or completely correct. For example, when a question asked a participant to provide the complete name of Botox (which is botulinium toxin), and the answer contained one of the words botulinium or toxin (but not both), the answer was scored as partially correct; when a participant would provide the response botoxicum, this answer was scored as a commission error. Two raters scored 17.5% (a total of 654) responses on questions about factual information and agreement was high (responses were scored using a nominal scale, Kappa = .96).

2.4.2. Diagram completion task

In line with the scoring of generated summaries by Thiede et al. (2003), the responses in the 4 filled-out text boxes in the completed diagrams were scored as correct (a correct step in the causal chain is provided) and factual information (the response refers to a detail from the text, but not to a step in the causal chain). For instance, when a person only typed "Indian Ocean" in a text box of the diagram about the Suez Canal, this response was scored as factual information, because the response contains a detail from the text but not the correct step in the causal chain. Furthermore, when the participants provided a response that did not come from the text, this was scored as commission error. When no response was provided in a diagram textbox, this was scored as omission.

Two raters independently coded 13.6% of the diagrams (a total of 400 text box responses) and inter-rater agreement was high (Kappa = .80).

2.5. Analyses

As in prior research on metacomprehension (Thiede et al., 2009), our focus was on relative accuracy. Relative accuracy is the degree to which the predictions discriminate between the different levels of performance on the criterion test for one text relative to another. We used the gamma correlation to measure relative accuracy (Nelson, 1984), because this non-parametric statistic has been considered one of the most appropriate measures of relative accuracy (Nelson, 1984), and it has been reported in prior research on the effects of generation tasks on judgment accuracy (e.g., Thiede et al., 2003, 2005). Although Nelson (1984) focused on the correlation between variables with only two levels (i.e., a 2×2 data array), gamma can be computed across variables with more than two levels (see Nelson, 1984, p. 124), which is common in the metacomprehension literature. The value of gamma indicates the strength of the association between POPs and test performance; the values range from -1 (indicating a perfect negative association) to $+1$ (indicating a perfect positive association). A value of zero indicates that there is no association between POPs and performance.

To assess monitoring accuracy for causal relations, intra-individual gamma correlations were calculated between the participants' POPs and their actual test scores for causal relations. For 12 participants we could not calculate monitoring accuracy for learning of causal relations; nine had invariance in POPs, and three participants' test responses on questions about causal relations were not saved by the computer.

To assess monitoring accuracy for factual information, intra-individual gamma correlations were calculated between POPs for learning of factual information and test scores for questions about factual information. Monitoring accuracy for learning of factual information could not be calculated for 14 participants; 12 had invariance in POPs, and two participants' responses to questions about factual information were not saved by the computer.

For regulation of study, the intra-individual gamma correlation between POPs for causal relations and whether a text was selected for restudy (yes = 1, no = 0) was calculated, as operationalized by Thiede et al. (2003). A correlation below 0 (i.e., negative) between POPs and restudy indicates that predictions are translated into the selection of less-well-known texts for restudy.

To obtain insight into cue diagnosticity, we investigated whether the learners' responses in the diagram text boxes were related to later test performance on questions about causal relations. As previously mentioned, responses in the diagram text boxes were scored as correct response, omission, commission error, or response containing only factual information but no causal relations from the text. Because the number of the responses and the number of correct relations are measured on an interval scale, Pearson correlations were calculated between the number of such responses (individually for different types of responses) per diagram and the number of correct relations that were provided at the later test. A correlation involving a given response type (i.e., cue) that is greater than 0 would indicate that the cue is diagnostic, with increasingly higher values (closer to +1.0) indicating greater diagnosticity.

Cue utilization was estimated by computing the relation between learners' POPs and the number of the response types provided during diagram completion. We calculated intra-individual gamma correlations between the number of the response types in the diagrams about each text and POPs about learning of causal relations for each text. A correlation involving a particular response type (i.e., cue) that is greater than 0 indicates would suggest the cue is used for making POPs, with increasingly higher correlations (closer to +1.0) indicating greater utilization.

3. Results

3.1. Monitoring accuracy

Fig. 2 depicts the effect of diagram-completion task on monitoring accuracy for learning of causal relations and factual information. The figure shows that for learning of causal relations, monitoring accuracy is higher for the immediate diagram group and, especially, for the delayed diagram group than for the no-diagram group. For learning of factual information, this pattern is not evident.

A mixed ANOVA was conducted to evaluate the effects of instructions on monitoring accuracy for the two types of information (causal relations and factual information). All effects are reported as significant

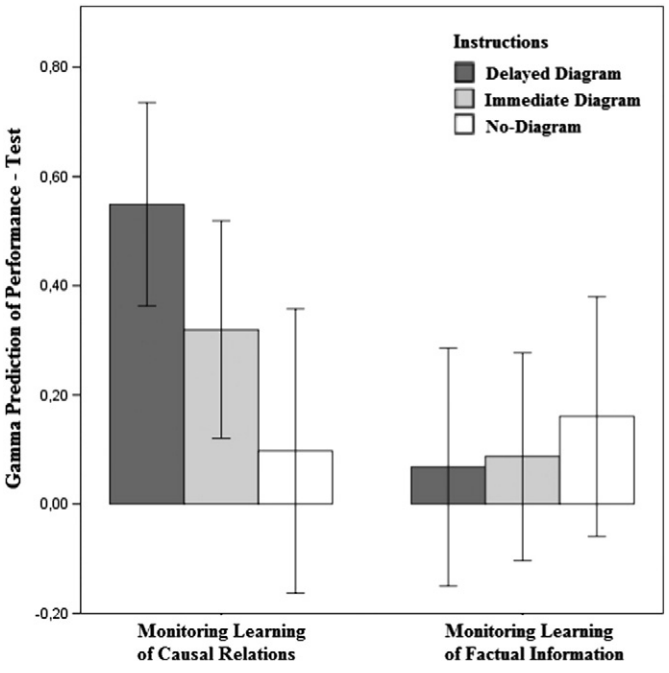

Fig. 2. Effects of instructions on monitoring accuracy for learning of causal relations and factual information. Error bars indicate the 95% confidence interval.

at $p < .05$. For significant effects, partial eta squared is presented as a measure of effect size.

First of all, a significant main effect of information type occurred, $F(1, 104) = 5.94$, $p = .016$, $\eta_p^2 = .054$, showing that overall monitoring accuracy was higher for learning of causal relations (M gamma = .21, $SD = .66$) than for learning of factual information (M gamma = .10, $SD = .61$). There was no significant main effect of instruction, indicating that overall monitoring was in general the same for the participants in the three groups, $F(2, 104) = 1.642$, $p = .199$. However, a significant interaction effect occurred between monitoring accuracy for the two types of information (causal relations and facts) and instruction, $F(2, 104) = 3.089$, $p = .049$, $\eta_p^2 = .056$. As shown in Fig. 2, this interaction effect shows that monitoring for learning of causal relations was affected by the diagram instructions, whereas there seems to be no effect of diagram completion on monitoring accuracy for learning of factual information.

To break down this interaction effect, a one-way ANOVA was conducted separately for learning of causal relations and for learning of factual information. The analysis for causal relations shows that instructions significantly affected monitoring accuracy, $F(2, 108) = 5.466$, $p = .005$, $\eta_p^2 = .092$. Bonferroni-corrected post-hoc tests show that monitoring accuracy was significantly higher for the delayed-diagram group (M gamma = .56, $SD = .57$) than for the no-diagram group ($M = .07$, $SD = .73$; $p = .004$). There were no significant differences in monitoring accuracy for learning of causal relations between the immediate diagram ($M = .28$, $SD = .62$) and the no-diagram group, $p = .495$, and between the immediate diagram and the delayed-diagram group, $p = .180$. Thus, Hypothesis 1, which is that monitoring accuracy will be greater when the participants complete diagrams than when they do not, was only partially confirmed: Immediate diagram completion did not significantly improve monitoring accuracy, whereas delayed diagram completion helped relative to not completing diagrams.

Hypothesis 2, stating that delayed diagrams would lead to the most accurate monitoring in comparison to the no-diagram and the immediate-diagram group was partially confirmed. Monitoring accuracy was significantly higher when diagrams were completed after a delay than when no diagrams where completed. Even though it was not significant, a trend occurred in the expected direction, with delayed diagram completion leading to more accurate monitoring in comparison to immediate diagram completion.

Next, consider the analysis of factual information. A one-way ANOVA revealed that the particular task did not affect monitoring accuracy in this case, $F(2, 106) = .37$, $MSE = .381$, $p = .692$, which indicates that diagram-completion tasks only affected monitoring for learning of causal relations (confirming Hypothesis 3).

3.2. Diagram responses and their contribution to monitoring accuracy

3.2.1. Diagram responses

We explored whether learners' responses in the diagrams were diagnostic of test performance (cue diagnosticity, Question 1a) and whether learners actually based their predictions on these cues (cue utilization, Question 1b). When monitoring learning, learners seem to be quite accurately calibrated for correct responses and omissions, whereas their calibration for incorrect responses (commission errors) is usually inaccurate (Lipko et al., 2009; Van Loon, de Bruin, van Gog & van Merrienboer, 2013b). Therefore, we specifically focused on cue diagnosticity and cue utilization of omissions and fully correct responses, because we would expect these diagram response categories to be most diagnostic, and most often appropriately utilized to predict performance.

Before we present the correlational analyses relevant to diagnosticity and utilization, however, we consider first the number of responses (correct relations, omissions, responses containing factual information, and commission errors) provided in the text boxes for the delayed and the immediate diagram groups, which are presented in Table 2. The participants who completed diagrams immediately after reading included more correct relations in the diagrams than those who completed diagrams at a delay, $t(81) = 1.99$, $p = .050$, Cohen's $d = .44$. Following the delayed diagram task, the participants made more omissions as compared to those in the immediate diagram group, $t(81) = 2.16$, $p = .034$, Cohen's $d = .47$. Differences between the immediate and delayed diagram group were not significant for the number of commission errors in diagrams ($p = .470$) or for the number of responses containing factual information ($p = .366$). Most importantly, inspection of Table 3 seems to confirm that omissions and correct responses in diagrams were diagnostic, and that these response types were utilized by the learners to make POPs. In the next sections, we present separate inferential analyses of cue diagnositicity and cue utilization, which strongly confirm these observations.

Table 2
Responses in diagrams.

	Correct relations	Omission	Factual information	Commission error
Immediate diagrams	2.34 (.61)	.41 (.43)	.47 (.24)	.67 (.44)
Delayed diagrams	2.06 (.67)	.67 (.63)	.42 (.27)	.73 (.43)

Note. Mean number of correct relations, omissions, responses containing factual information, and commission errors in the four text boxes for the immediate diagram and delayed diagram groups. Standard deviations of the mean in parentheses.

3.2.2. Cue diagnosticity

Table 3 includes the Pearson correlations between the number of response types in the diagram text boxes and the number of correct relations at the test, indicating cue diagnosticity of the diagram responses. This table shows that diagram responses were diagnostic of later test performance (most correlations were significantly different from 0). The correlations involving correct responses were highly positive (.49 and .53 for delayed and immediate diagrams, respectively), indicating that providing a correct response in the diagram was strongly related to providing correct relations at the later test. Moreover, the correlation involving omissions was negative for both the delayed (−.39) and the immediate (−.24) diagram groups, indicating that when learners provided fewer responses in the diagram about a text, they were less likely to produce correct causal relations at the later test.

A mixed ANOVA was used to investigate the effects of the immediate and delayed diagram completion on cue diagnosticity for omissions and correct relations in the diagrams. The relation between the number of omissions and test score was negated for the analysis, in order to use

Table 3
Cue diagnosticity and cue utilization.

	Cue diagnosticity	Cue utilization
Correct relations		
Delayed diagrams	.49[a] (.34)	.59 (.42)[a]
Immediate diagrams	.53[a] (.41)	.23 (.60)[a]
Omissions		
Delayed diagrams	−.39[a] (.36)	−.64 (.58)[a]
Immediate diagrams	−.24[a] (.39)	−.50 (.64)[a]
Commission errors		
Delayed diagrams	−.17[a] (.41)	−.16 (.66)
Immediate diagrams	−.30[a] (.41)	.00 (.64)
Factual information		
Delayed diagrams	−.09 (.44)	.20 (.79)
Immediate diagrams	−.22[a] (.43)	−.11 (.65)

Note. Cue diagnosticity is calculated by the Pearson correlation between diagram responses and test scores on questions about causal relations. Cue utilization is calculated by the Gamma correlation between diagram responses and POPs about learning of causal relations. The table presents the means of the intra-individual gamma correlations. Standard deviations of the mean in parentheses.

[a] Correlation significantly differs from zero, $p < .01$.

positive values, so that the strength of the correlation for the two response types can be assessed and compared. Thus, if one of the diagram instructions would lead to higher cue diagnosticity, we would expect to find an effect of the timing of diagram completion (delayed, immediate).

Table 3 shows cue diagnosticity of omissions and correct relations as an effect of the timing of the diagram completion. There was no significant effect of instructions, $F(1, 81) = .123$, $p = .727$. Thus, completing delayed diagrams did not provide learners with more diagnostic cues than did completing immediate diagram completion.

In the analysis of cue diagnosticity described above, the cues did not necessarily need to correspond to the actual outcomes of the test; for instance, overall correct diagram completion was predictive of test performance, so it is considered a diagnostic cue. Even so, a participant may have correctly generated one relationship during diagram completion for a particular text, but when tested later, the participant may have been incorrect about that relationship but correctly responded about another one instead. In this case, the cue (recalling one relationship) was not diagnostic of eventually responding correctly about that relationship again. Accordingly, we also evaluated how well correct responses during diagram completion corresponded to responding correctly about the same information on the test. In particular, we computed the percentage of correct responses during diagram completion that were also correct on the final test; a higher value indicates higher correspondence. The percentage of correct responses that were both correct during diagram completion and on the final test was 87.60% ($SD = 18.28$) for the delayed diagram group and 87.17% ($SD = 18.18$) for the immediate diagram group. The groups did not differ, $p = .81$, and most importantly, the correspondence between correct diagram responses and test responses was high.

3.2.3. Cue utilization

The results in Table 3 indicate that following both delayed and immediate diagram completion, the learners seem to use the number of omissions and correct relations in their completed diagrams as a cue for POPs. They did not seem to use their commission errors and the responses containing factual information in the diagrams as a cue when providing their POPs (i.e., gamma correlations were not different from zero for these response types). Mixed ANOVA was conducted to investigate possible differences between the groups in cue utilization. First of all, this analysis shows a significant main effect of the response type in the diagrams on POPs, $F(1, 61) = 68.54$, $p < .001$, $\eta_p^2 = .529$. The correlation between diagram responses and POPs was negatively related to the number of omissions in the diagrams (indicating that the participants provided lower POPs when their diagram text boxes contained more omissions), and positively related to the number of correct responses in the diagrams (indicating that participants provided higher POPs when their diagram text boxes contained more correct relations). The correlation between omissions in diagrams and POPs was negated, in order to compare the strength of the relation for the two response types. There was a significant interaction effect between cue-utilization and the timing of the diagram completion, $F(1, 61) = 7.306$, $p = .009$, $\eta_p^2 = .107$. As evident from Table 3, omissions in diagrams were also used as cue for the POPs, with the correlation between the number of omissions and POPs being strong for both groups. Follow-up t-tests with a Bonferroni correction showed no significant difference between the immediate and the delayed diagram group in cue utilization of omissions, $t(61) = .79$, $p = .864$. However, a difference occurred between the delayed and the immediate diagram group in cue utilization of the number of correct relations, $t(76) = 3.07$, $p = .006$, Cohen's $d = .69$. Thus, delaying diagrams led learners to utilize the number of correct relations more when making their POPs.

3.3. Regulation

The participants selected 48.75% ($SD = 27.04$) of texts for restudy following delayed diagram completion, whereas they only selected 33.3% ($SD = 26.11$) following immediate diagrams and 36.69% ($SD = 25.41$) following the no-diagram instructions. The percentage of texts selected for restudy was significantly affected by the diagram task, $F(2, 120) = 3.52$, $p = .033$, $\eta_p^2 = .055$. Following delayed diagrams, the participants selected more texts for restudy than following immediate diagrams ($p = .039$). There were no significant differences between the delayed diagram and the no-diagram group ($p = .073$) and between the immediate diagram and the no-diagram group ($p > .999$).

Fig. 3 indicates the relation between POPs for causal relations and the restudy selections, as well as POPs for factual information, and the restudy selections for the three groups. This figure shows that gamma correlations between POPs and selections for restudy were highly negative for all groups, and for both types of information. These gamma correlations indicate a strong relation between monitoring judgments and restudy selections. With a mixed ANOVA we evaluated whether restudy selections were more related to POPs for learning of causal relations than to POPs for learning of factual information, and whether instructions differentially affected restudy selections. There was no main effect of the information type on the correlations between POP and restudy, $F(1, 88) = .23$, $p = .634$. Further, there was no main effect of instruction, $F(2, 88) = .016$, $p = .984$. Interestingly, there was a significant interaction effect between instruction and information type, $F(2, 88) = 4.80$, $p = .011$, $\eta_p^2 = .098$. As visible in Fig. 3, this interaction effect indicates that restudy selections were more related to POPs for learning of causal relations in the delayed-diagram group and the control group, whereas restudy selections for the immediate diagram group were more strongly related to POPs for learning of factual information.

Means across individual gamma correlations between POPs for causal relations and selection for restudy were −.74 ($SD = .54$) for delayed diagrams, −.53 ($SD = .71$) for immediate diagrams, and −.69 ($SD = .53$) for the no-diagram group. Mean gamma correlations between POPs for factual information and selection for restudy were −.57 ($SD = .71$) for delayed diagrams, −.75 ($SD = .35$) for immediate diagrams, and −.61 ($SD = .59$) for the no-diagram group. These gamma correlations indicate a strong relation between POPs and restudy selections. There was no significant effect of diagram completion group on the correlations between

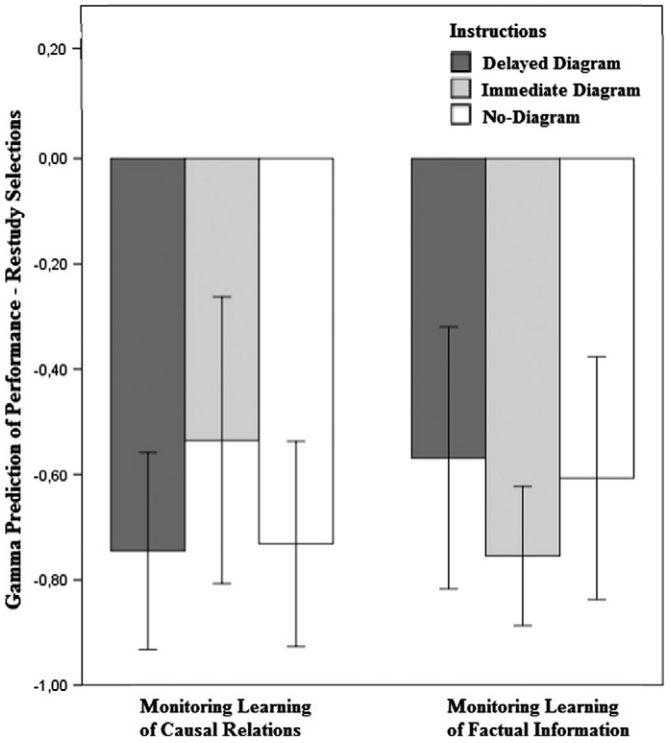

Fig. 3. Effects of instructions on regulation of study. Error bars indicate the 95% confidence interval.

POPs for causal relations and restudy selections, $F(2, 92) = 1.03$, $p = .362$, and no significant effect of diagram completion group on the correlation between POPs for factual information and restudy selections, $F(2, 88) = .86$, $p = .423$. These findings show that, even though monitoring for learning of causal relations was less accurate in the immediate diagram group and the no-diagram group, restudy selections were still highly related to their POPs, which is consistent with Hypothesis 4.

3.4. Supplementary analyses

In this section, we briefly discuss analyses relevant to whether the diagram-completion tasks influence POP magnitude and test performance. Bonferroni corrections were used for the post-hoc tests. Although these measures were not critical for evaluating our key predictions, we report them here for completeness.

3.4.1. POP magnitude

Mean POPs for causal relations and factual information are presented in Table 4. The diagram-completion group significantly affected the POP level for learning of causal relations, $F(2, 120) = 3.97$, $p = .021$, $\eta_p^2 = .062$. Following the immediate diagram task, POPs for causal relations (57.35) were significantly higher than POPs following the no-diagram comparison task (46.76), $p = .022$. For the delayed diagram group, causal relation POPs (54.32) did not significantly differ from the immediate diagram group ($p > .999$) and the no-diagram group ($p = .144$).

The diagram-completion group also had an effect on the level of POPs for factual information, $F(2, 120) = 3.64$, $p = .029$, $\eta_p^2 = .057$. The mean level of POPs was higher for the delayed diagram group ($M = 57.35$) than for the no-diagram group ($M = 47.70$, $p = .033$). The immediate diagram group ($M = 55.34$) did not significantly differ from the delayed diagram ($p > .999$) and the no-diagram group ($p = .144$).

3.4.2. Test performance and study time

Table 4 includes the mean percentage of correct performance for questions about causal relations (test performance for questions about causal relations could range from 0 to 4) and the mean percentage of correct responses to questions about factual information (test performance for questions about factual information could range from 0 to 5). Instructions significantly affected performance for questions about causal relations, $F(2, 119) = 9.28$, $p < .001$, $\eta_p^2 = .14$. Performance was highest for the immediate diagram group ($M = 2.40$) as compared to the delayed diagram ($M = 1.97$, $p = .004$) and the no-diagram group ($M = 1.77$, $p < .001$).

Mean study time per text (refer to Table 4) was also significantly affected by group, $F(2, 120) = 5.65$, $p = .004$, $\eta_p^2 = .086$. Study time per text was longer for the immediate diagram group ($M = 97.71$ s) when compared to the delayed diagram group ($M = 81.47$ s, $p = .014$) and the no-diagram group ($M = 80.51$ s, $p = .011$). There was no significant difference in test performance between the delayed diagram group and no-diagram group ($p = .175$), and there was no significant difference in study time for these two groups, $p > .999$. In addition, group did not affect test performance on questions about factual information, $F(2, 120) = .44$, $p = .647$.

4. Discussion

This study demonstrates that adolescents' monitoring accuracy for learning causal relations from texts improved when they completed diagrams prior to predicting performance, in comparison to learners who did not use causal diagrams. Moreover, the timing of the diagram completion task was an important factor when aiming to improve monitoring accuracy for learning of causal relations. The participants who completed the diagrams after a delay showed most accurate monitoring in comparison to the control group. Furthermore, the findings imply that the diagram completion task focused learners on understanding of cause-and-effect relations presented in the studied text, and not on learning of text base factual information.

These findings are consistent with research demonstrating beneficial effects of delayed generation tasks, such as summary generation (Thiede & Anderson, 2003), keyword generation (De Bruin et al., 2011; Thiede et al., 2003) and sentence generation (Van Loon, De Bruin, Van Gog, & Van Merriënboer, 2013a) on monitoring accuracy. The present results extend previous findings by demonstrating that a diagram completion task can focus learners specifically on learning of complex cause-and-effect relations. This study is the first to show that instructional strategies can specifically improve monitoring for complex higher-order ideas in text. A further contribution of this research is that the study was conducted with adolescents in their classroom context. Even though a comparable level of monitoring accuracy has been demonstrated with adults (e.g. Thiede et al., 2003), this is the first study showing a relatively high level of monitoring accuracy (gamma = .56) in adolescent learners when reading comprehension of complex information. Note that De Bruin et al. (2011) improved the monitoring accuracy for younger learners when keywords were generated after reading expository texts, their gamma correlations after delayed generation were .27 (Experiment 1; seventh graders) and .42 (Experiment 2; sixth graders).

Even though monitoring accuracy was higher than zero for the immediate diagram group, only the delayed diagram group showed significantly more accurate monitoring for learning of causal relations than the no-diagram control group. The experimental design allowed us to investigate why delayed diagram completion had most beneficial effect on monitoring accuracy. Analyses showed that both the immediate diagram and the delayed diagram completion tasks established cues that were diagnostic of learners' future test performance. Namely, for both diagram groups, within-participant variability in completing diagram boxes (i.e. correct responses) and in not completing boxes (i.e., omission errors) predicted final test performance (Question 1a). Importantly, monitoring judgments can only be accurate when learners utilize cues that are diagnostic of their actual learning (Brunswik, 1956; Koriat, 1997). Findings indicated that following delayed diagram completion, participants were more successful at utilizing diagnostic cues; their POPs were more strongly related to the diagnostic cues (the number of correct relations generated and the number of omissions) that were established during diagram completion.

From these findings, the question arises as to why delayed diagram completion leads to more successful utilization of diagnostic cues. A possible explanation for our finding that delayed diagram completion supported higher judgment accuracy and led to the most effective cue utilization is that there was more variability in correct responses in

Table 4
POPs, study time, and test performance.

Instruction	POP causal relations	POP facts	Study time per text	Percentage correct causal relations	Percentage correct facts
Delayed diagrams	54.32 (17.24)	57.35 (16.66)	81.47 (26.57)	49.22 (17.05)	32.57 (14.06)
Immediate diagrams	57.35 (18.83)	55.34 (18.32)	97.71 (29.29)	59.98 (16.08)	30.14 (11.77)
Control	46.76 (15.27)	47.70 (15.87)	80.51 (20.43)	44.19 (16.67)	30.72 (12.71)

Note. Statistics are presented for the instructions delayed diagram completion; immediate diagram completion; and control group. The table presents mean POPs for causal relations (ranging from 0 to 100%); mean POPs for factual information (ranging from 0 to 100%); study time per text (in seconds); percentage of correct causal relations at the test; and the percentage of correct responses to questions about factual information. Standard deviations of the mean in parentheses.

the diagrams following delayed diagram completion than following immediate diagram completion. If there was more variability in this diagnostic cue, learners might have found it easier to use it, perhaps because it was more salient to them. If so, the intra-individual standard deviations for the correct responses in the diagrams should then be higher for the delayed diagrams than the immediate diagrams. We evaluated this possibility and found that the intra-individual standard deviations for the mean number of correct causal relations in the generated diagrams did not differ between the immediate and the delayed diagram groups (mean $SD = .30$ for immediate diagrams; mean $SD = .28$ for delayed diagrams, $t(81) = .38$, $p = .705$).

Given that differences in cue variability do not seem to explain why learners more effectively used diagnostic cues when completing diagrams after a delay rather than immediately, it is an open issue why this effect occurred. Future research should investigate this issue. One possibility is to ask learners to think aloud during diagram completion and when predicting performance, which could provide insight into the effects of delaying diagram completion on cue utilization.

Note that completing diagrams immediately during learning led to higher test scores when learners were tested for learning of causal relations. This finding is in line with research demonstrating that interaction with diagrams during text study can support deep processing, and therefore improve deep comprehension (Cromley et al., 2013). Our findings that reading times were longer for the immediate diagram group, and that they provided more correct relations in their diagrams support the notion that immediate diagram completion focused learners on deeper processing. The observation that reading times were longer for the group who completed immediate diagrams than for the delayed diagram group and the no-diagram group is consistent with the findings reported by Ainsworth, Prain, and Tytler (2011), who showed that interaction with diagrams during reading led to more engagement with the learning task.

Note that learners could study each text only once, and that they had no opportunity to return to the text after reading, even though they indicated which texts they would like to restudy if they had the chance. Our findings show that, regardless of group and the level of monitoring accuracy, restudy selections were based on participants' subjective judgments of how well they had learned the texts. Thus, even though in the present study the POPs of the immediate diagram and the no diagram group were not highly accurate, learners still used their POPs for their study selections. Importantly, research by Thiede and Anderson (2003) showed that honoring restudy choices drastically improved performance of the delayed generation group, in comparison to an immediate generation group and a control group. Using the more accurate predictions to make restudy decisions after delayed diagram completion would be expected to support more effective self-regulated learning (Thiede, 1999). Therefore, if the participants would have had the chance to implement their restudy choices, this would most likely allow the delayed diagram group to surpass test performance about comprehension of causal relations in comparison to the immediate diagram group and the no-diagram group. Furthermore, the group completing delayed diagrams did not only select more effectively because they based restudy selections on more accurate POPs; they also selected more texts for restudy in comparison to the immediate diagram group. This might imply that learners in the delayed diagram group were more motivated to restudy the texts. The findings indicate that the attempt to complete a diagram provided learners with diagnostic cues, regardless of whether the diagrams were completed immediately or after a delay. Possibly, learners selected more texts for restudy when diagrams were completed after a delay because they realized that they already forgot some previously studied information.

Our evidence suggests that in a self-regulated learning context, delayed diagram completion may emerge as the more effective task to improve self-regulated learning of texts containing cause-and-effect relations. Future research could further investigate the effect of honoring restudy selections, and could address the effects of delayed generation tasks on learners' motivation to restudy information.

Interestingly, the restudy selections seemed to be more related to POPs for learning of causal relations for both the delayed diagram and the control group, whereas restudy selections for the immediate diagram group were more related to POPs for factual information. This might indicate that the immediate diagram group focused learners more on their detail learning, which seems in line with findings by Anderson and Thiede (2008). They showed that participants focused more on details when generating immediate summaries, whereas the participants in the delayed summary group focused more on deeper comprehension of gist. However, because monitoring for learning of factual information was inaccurate, it seems unlikely that the immediate diagram group would have improved their learning of factual information through restudy if the study selections had been honored.

The level of POPs was higher following diagram completion than for the control group, which might imply that interaction with diagrams led to a higher level of confidence. However, this could also indicate that learners used the POP scale differently after interacting with diagrams. The test for learning of factual information seemed more difficult than the test for learning of cause-and-effect relations. Even though learning of facts was weaker than learning of causal relations in all three groups, the magnitude of their POPs imply that learners did not notice this; they seemed to be more overconfident for their learning of factual information than for their learning of cause-and-effect relations. This might have occurred because in the instructional phase prior to the experiment, more attention was paid to the causal relations and the structure of cause-and-effect relations in texts than to the factual information in the text. Therefore, learners might have focused more on the causal relations than on the factual information during study and hence performed better. Thus, the finding that learners were more overconfident for their learning of factual information has two possible (non-mutually exclusive) explanations. Lack of knowledge is often related to overconfidence (e.g., Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008; Lipko et al., 2009; Van Loon, de Bruin, van Gog & van Merrienboer, 2013b), and the learners may have had more difficulty monitoring and evaluating how well they had learned the factual knowledge. Alternatively, the learners may have anchored their judgments near the middle of the scale (e.g., because they could not predict overall test difficulty) and hence used the middle of the scale to indicate a general lack of confidence (e.g., Dunlosky, Serra, Matvey, & Rawson, 2010; Keren, 1991). If so, the similar judgment magnitudes for the two kinds of materials (Table 4) would indicate that the learners were equally unaware of what their absolute level of performance would be for both kinds of questions, and hence the greater overconfidence for factual questions would be an artifact of differences in performance (e.g., Connor, Dunlosky, & Hertzog, 1997; Krueger & Mueller, 2002). Competitively evaluating these possibilities is an important avenue for future research.

The finding that delayed diagram completion is a promising task to improve monitoring accuracy when studying texts containing causal relations provides a novel contribution for educational practice. The present study also highlights a novel application of tasks asking learners to interact with diagrams (e.g., Cromley et al., 2013). This study was the first to show that adolescents' monitoring accuracy for higher-order ideas in text can be improved in an educational setting. For optimal comprehension of causal relations, the learners should study the texts, interact with diagrams after a delay, and then use their monitoring to further restudy texts of which the causal relations have been least well learned.

Acknowledgments

This research was supported by the Netherlands Organization for Scientific Research (NWO), grant 411.07.151. The authors would like

to thank Gerrit Drost who assisted with the programming of the experimental tasks and Michael Mueller for comments on an earlier draft of this manuscript.

Appendix A

Text "The Suez Canal"

"The Suez Canal, which connects the Indian Ocean and the Mediterranean Sea with each other, is of great importance to the world. Originally, there was no natural water connection between the Atlantic and the Indian Ocean. Between these two seas is a desert. This meant that trading ships that traveled from the harbor city Jeddah in Saudi Arabia to Europe had to make a long journey around the whole African continent. It was therefore decided that a shorter waterway was needed that would connect the two oceans with each other. For this reason, the Suez Canal, which was designed by the Austrian engineer Alois Negrelli, was dug. For years, workers were digging; the canal was finally opened in 1869 for shipping. By the digging of the Suez Canal, the distance from the harbor city of Jeddah to the harbor city of Rotterdam has been reduced by 40%. Through the Suez Canal, the distance between these cities is 6,337 nautical miles, when ships sail around the African continent this distance is 10,743 nautical miles."

Text "Botox"

Botox is the abbreviation of Botulinium Toxin, this is a poison that is produced by the bacterium *Clostridium botulinum*. This substance blocks the signal between the nerves and the muscles in the skin. Since 1989, use of Botox is permitted, although this is strictly controlled in The Netherlands. In 2004, 28 people died in America, they had an accident with an incorrect dosage of Botox. Due to the blocking of the signal between the nerves and skin, originally, Botox was particularly used against muscle contractions, for example with patients who could not control muscle contractions and continuously blinked their eyes. By injecting Botox around the eyes, the muscles are paralyzed and the muscle contractions disappear. Because Botox blocks the signal between the nerves and the muscles in the skin, this is also used in plastic surgery to smoothen the skin: It can reduce the wrinkles around the eyes and the forehead. Because wrinkles are reduced, this treatment makes people look younger. The effect of such a treatment usually lasts between 1 and 6 months. However, this treatment against wrinkles between the eyes and on the forehead can also undesirably change peoples' face expressions.

Appendix B

Example questions:
Questions about causal relations:

- The distance for trading ships that sail between Jeddah and Rotterdam has been reduced a lot. For what reasons has the distance between Jeddah and Rotterdam been reduced?
- Botox blocks the signal between the nerves and the skin. What are the effects of this?

Questions about factual information:

- In what year was the Suez Canal opened for ships?
- From which country was the engineer who designed the Suez Canal?
- What is the full name of Botox?
- Since when has use of Botox been officially permitted?

References

Ackerman, R., & Goldsmith, M. (2011). Metacognitive regulation of text learning: On screen versus on paper. *Journal of Experimental Psychology: Applied*, *17*, 18–32. http://dx.doi.org/10.103/a0022086.

Ackerman, R., Leiser, D., & Shpigelman, M. (2013). Is comprehension of problem solutions resistant to misleading heuristic cues? *Acta Psychologica*, *143*, 105–112. http://dx.doi.org/10.1016/j.actpsy.2013.02.004.

Ainsworth, S., Prain, V., & Tytler, R. (2011). Drawing to learn in science. *Science*, *333*, 1096–1097. http://dx.doi.org/10.1126/science.1204153.

Alvermann, D. E. (2002). Effective literacy instruction for adolescents. *Journal of Literacy Research*, *34*(2), 189–208. http://dx.doi.org/10.1207/s15548430jlr3402_4 (Anderson, M. C. M., &).

Anderson, M. C. M., & Thiede, K. W. (2008). Why do delayed summaries improve metacomprehension accuracy? *Acta Psychologica*, *128*, 110–118. http://dx.doi.org/10.1016/j.actpsy.2007.10.006.

Brunswik, E. (1956). Perception and the representative design of psychological experiments. Berkeley: University of California Press.

Butcher, K. R. (2006). Learning from text with diagrams: Promoting mental model development and inference generation. *Journal of Educational Psychology*, *98*, 182–197. http://dx.doi.org/10.1037/0022-0663.98.1.182.

Connor, L. T., Dunlosky, J., & Hertzog, C. (1997). Age-related differences in absolute but not relative metamemory accuracy. *Psychology and Aging*, *12*(1), 50–71. http://dx.doi.org/10.1037//0882-7974.12.1.50.

Cromley, J. G., Bergey, B. W., Fitzhugh, S., Newcombe, N., Wills, T. W., Shipley, T. F., et al. (2013). Effects of three diagram instruction methods on transfer of diagram comprehension skills: The critical role of inference while learning. *Learning and Instruction*, *26*, 45–58. http://dx.doi.org/10.1016/j.learninstruc.2013.01.003.

Cromley, J. G., Snyder-Hogan, L. E., & Luciw-Dubas, U. A. (2010). Reading comprehension of scientific text: A domain-specific test of the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology*, *102*, 687–700. http://dx.doi.org/10.1037/A0019452.

De Bruin, A.B. H., Thiede, K. W., Camp, G., & Redford, J. (2011). Generating keywords improves metacomprehension and self-regulation in elementary and middle school children. *Journal of Experimental Child Psychology*, *109*, 294–310. http://dx.doi.org/10.1016/j.jecp.2011.02.005.

De Bruin, A.B. H., & Van Gog, T. (2012). Improving self-monitoring and self-regulation: From cognitive psychology to the classroom. *Learning and Instruction*, *22*(4), 245–252. http://dx.doi.org/10.1016/j.learninstruc.2012.01.003.

Dunlosky, J., & Ariel, R. (2011). Self-regulated learning and the allocation of study time. In B. Ross (Ed.), *Psychology of Learning and Motivation*, *54*. (pp. 103–140). http://dx.doi.org/10.1016/B978-0-12-385527-5.00004-8.

Dunlosky, J., & Lipko, A.R. (2007). Metacomprehension: A brief history and how to improve it's accuracy. *Current Directions in Psychological Science*, *16*, 228–232. http://dx.doi.org/10.1111/j.1467-8721.2007.00509.x.

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self-evaluations undermine students' learning and retention. *Learning and Instruction*, *22*(4), 271–280. http://dx.doi.org/10.1016/j.learninstruc.2011.08.003.

Dunlosky, J., Serra, M., Matvey, G., & Rawson, K. A. (2010). Second-order judgments about judgments of learning. *Journal of General Psychology*, *132*, 335–346. http://dx.doi.org/10.3200/GENP.132.4.335-346.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes*, *105*, 98–121. http://dx.doi.org/10.1016/j.obhdp.2007.05.002.

Fry, A. F., & Hale, S. (1996). Processing speed, working memory, and fluid intelligence: Evidence for a developmental cascade. *Psychological Science*, *7*, 237–241. http://dx.doi.org/10.1111/j.1467-9280.1996.tb00366.x.

Gobert, J.D., & Clement, J. J. (1999). Effects of student-generated diagrams versus student-generated summaries on conceptual understanding of causal and dynamic knowledge in plate tectonics. *Journal of Research in Science Teaching*, *36*(1), 39–53. http://dx.doi.org/10.1002/(sici)1098-2736(199901)36:1<39::aid-tea4>3.0.co;2-i.

Graesser, A.C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, *101*, 371–395. http://dx.doi.org/10.1037/0033-295x.101.3.371.

Kail, R., & Salthouse, T. A. (1994). Processing speed as a mental capacity. *Acta Psychologica*, *86*, 199–225. http://dx.doi.org/10.1016/0001-6918(94)90003-5.

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica*, *77*, 217–273. http://dx.doi.org/10.1016/0001-6918(91)90036-Y.

Kimball, D. R., & Metcalfe, J. (2003). Delaying judgments of learning affects memory, not metamemory. *Memory & Cognition*, *31*(6), 918–929. http://dx.doi.org/10.3758/Bf03196445.

Kintsch, W. (1998). Comprehension: A paradigm for cognition. Cambridge: University Press.

Kintsch, W., Welsch, D.M., Schmalhofer, F., & Zimny, S. (1990). Sentence memory: A theoretical analysis. *Journal of Memory and Language*, *29*, 133–159. http://dx.doi.org/10.1016/0749-596x(90)90069-C.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*, 349–370. http://dx.doi.org/10.1037/0096-3445.126.4.349.

Koriat, A. (2012). The relationships between monitoring, regulation and performance. *Learning and Instruction*, *22*(4), 296–298. http://dx.doi.org/10.1016/j.learninstruc.2012.01.002.

Koriat, A., Ackerman, R., Adiv, S., Lockl, K., & Schneider, W. (2014). The effects of goal-driven and data-driven regulation on metacognitive monitoring during learning: A developmental perspective. *Journal of Experimental Psychology: General*, *143*, 386–403. http://dx.doi.org/10.1037/a0031768.

Krebs, S. S., & Roebers, C. M. (2010). Children's strategic regulation, metacognitive monitoring, and control processes during test taking. *British Journal of Educational Psychology*, *80*, 325–340. http://dx.doi.org/10.1348/000709910X485719.

Krebs, S. S., & Roebers, C. M. (2011). The impact of retrieval processes, age, generalachievement level, and test scoring scheme for children's metacognitive

monitoring and controlling. *Metacognition and Learning*. http://dx.doi.org/10.1007/s11409-011-9079-3.
Krueger, J., & Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology*, *82*, 180–188. http://dx.doi.org/10.1037//0022-3514.82.2.180.
Lipko, A.R., Dunlosky, J., Hartwig, M. K., Rawson, K. A., Swan, D., & Cook, D. (2009). Using standards to improve middle school students' accuracy at evaluating the quality of their recall. *Journal of Experimental Psychology: Applied*, *15*, 307–318. http://dx.doi.org/10.1037/a0017599.
Maki, R. H. (1998). Predicting performance on text: Delayed versus immediate predictions and tests. *Memory & Cognition*, *26*, 959–964. http://dx.doi.org/10.3758/bf03201176.
Mayer, R. E. (2003). The promise of multimedia learning: Using the same instructional design methods across different media. *Learning and Instruction*, *13*, 125–139. http://dx.doi.org/10.1016/S0959-4752(02)00016-6.
McCrudden, M. T., Magliano, J. P., & Schraw, G. (2011). The effect of diagrams on online reading processes and memory. *Discourse Processes*, *49*, 69–92. http://dx.doi.org/10.1080/01638531003694561.
McCrudden, M. T., Schraw, G., Lehman, S., & Poliquin, A. (2007). The effects of causal diagrams on text learning. *Contemporary Educational Psychology*, *32*, 367–388. http://dx.doi.org/10.1016/j.cedpsych.2005.11.002.
Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review*, *15*, 174–179. http://dx.doi.org/10.3758/pbr.15.1.174.
Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*, 109–133. http://dx.doi.org/10.1037/0033-2909.95.1.109.
Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect". *Psychological Science*, *2*, 267–270. http://dx.doi.org/10.1111/j.1467-9280.1991.tb00147.x.
Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation*, *26*, 125–141. http://dx.doi.org/10.1016/S0079-7421(08)60053-5.
Otero, J., Leon, J. A., & Graesser, A.C. (Eds.). (2002). *The psychology of science text comprehension*. Erlbaum: Mahwah NJ.
Ozuru, Y., Kurby, C., & McNamara, D. (2012). The effect of a metacomprehension judgment task on comprehension monitoring and metacognitive accuracy. *Metacognition and Learning*, *7*, 113–131. http://dx.doi.org/10.1007/s11409-012-9087-y.
Rawson, K. A., Dunlosky, J., & Thiede, K. W. (2000). The rereading effect: Metacomprehension accuracy improves across reading trials. *Memory & Cognition*, *28*, 1004–1010. http://dx.doi.org/10.3758/BF03209348.
Redford, J. S., Thiede, K. W., Wiley, J., & Griffin, T. D. (2012). Concept mapping improves metacomprehension accuracy among 7th graders. *Learning and Instruction*, *22*, 262–270. http://dx.doi.org/10.1016/j.learninstruc.2011.10.007.
Roebers, C. M., Von der Linden, N., & Howie, P. (2007). Favourable and unfavourable conditions for children's confidence judgments. *British Journal of Developmental Psychology*, *25*, 109–134. http://dx.doi.org/10.1348/026151006X104392.
Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 204–221. http://dx.doi.org/10.1037/0278-7393.26.1.204.
Thiede, K. W. (1999). The importance of accurate monitoring and effective self-regulation during multitrial learning. *Psychonomic Bulletin & Review*, *6*, 662–667. http://dx.doi.org/10.3758/BF03212976.
Thiede, K. W., & Anderson, M. C. M. (2003). Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology*, *28*, 129–160. http://dx.doi.org/10.1016/S0361-476x(02)00011-5.
Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of text. *Journal of Educational Psychology*, *95*, 66–73. http://dx.doi.org/10.1037/0022-0663.95.1.66.
Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *25*, 1024–1037. http://dx.doi.org/10.1037/0278-7393.25.4.1024.
Thiede, K. W., Dunlosky, J., Griffin, T. D., & Wiley, J. (2005). Understanding the delayed keyword effect on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 1267–1280. http://dx.doi.org/10.1037/0278-7393.31.6.1267.
Thiede, K. W., Griffin, T. D., Wiley, J., & Redford, J. (2009). Metacognitive monitoring during and after reading. In D. J. Hacker, J. Dunlosky, & A.C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 85–106). Routledge.
Van den Broek, P., Rapp, D. N., & Kendeou, P. (2005). Integrating memory-based and constructionist processes in accounts of reading comprehension. *Discourse Processes*, *39*, 299–316. http://dx.doi.org/10.1080/0163853X.2005.9651685.
Van Loon, M. H., De Bruin, A.B. H., Van Gog, T., & Van Merriënboer, J. J. G. (2013a). The effect of delayed-JOLs and sentence generation on children's monitoring accuracy and regulation of idiom study. *Metacognition and Learning*, *8*, 173–191. http://dx.doi.org/10.1007/s11409-013-9100-0.
Van Loon, M. H., de Bruin, A. B. H., van Gog, T., & van Merrienboer, J. J. G. (2013b). *Activation of inaccurate prior knowledge affects primary-school students' metacognitive judgments and calibration Learning and Instruction*, *24*, 152–510 http://dx.doi.org/1016/j.learninstruc.2012.08.005.